Week 21 : Point and Interval Estimators

pending tasks	
type	

Point and Interval Estimators

Estimators are used to deduce the population parameters using some sample data.

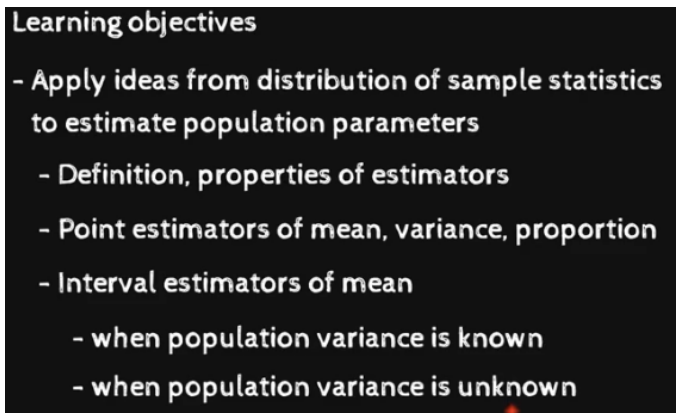


Fig.1 Module objectives.

Distribution of statistics over many samples is only studied in theory but in practice, parameters are inferred from a single sample.

Examples to Solve

- Given a sample data, there should be a way to generalize the observations made to a population.
- How to deal with the error in logging the data e.g. if the error of weighing scale is known, what can be said about the average weight of a person, observed over a week.
- Error margin perhaps decreases with an in increase in sample size. It depends on the sample size and how skewed the data is e.g. outcomes of survey data.
- A general template would be, given some sample data with the goal to make an inference about the population:
    1. Compute sample statistics
    2. Use knowledge of distribution of sample statistics
    3. Estimate population parameters

What are the Estimators

def estimator - (t = 0:33)

- An estimator is a statistic of a given sample. The value of the estimator, called an estimate, is used to predict a population parameter. Thus, estimator is a rule that is applied on the sample (e.g. sample mean) to get a concrete value called an estimate, which can be used to predict the population parameter.
- A point estimate is a single estimated value of the population from sample data.

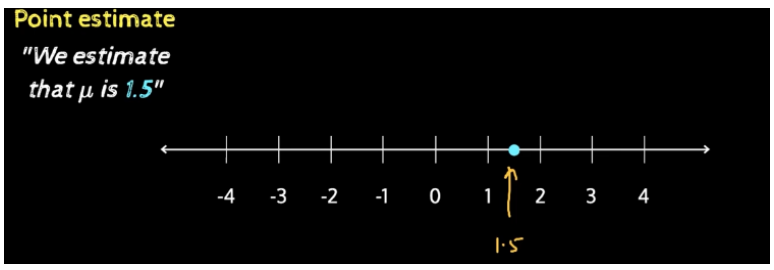


Fig.2a Example of point estimate.

- An interval estimate is an interval deduced from the sample data in which the population parameter might occur with certain confidence.

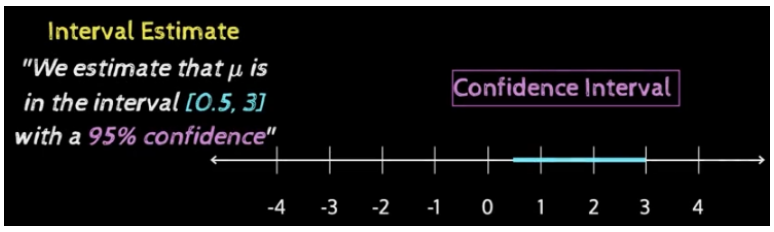


fig.2b Example of confidence interval.

Intervals are useful when narrow. A confidence intervals implies, using the example in fig.2b, if 100 samples were taken then in 95 samples the parameter would lie in the interval. The desired confidence and interval width depend on the application.

Properties of Estimator

def unbiased estimator - (t = 0:42)

def consistent estimator - (t = 3:13)

- An estimator is **unbiased** if the expected value of the estimator is equal to the parameter being estimated. E.g. mean ($E[\bar{X}] = \mu$), proportion ($E[\hat{p}] = p$)and variance ($E[s^2_{n-1}] = \sigma^2$) are simple unbiased estimators.

  This implies that the estimate can be lesser, greater or equal to the parameter but when taken for a large number of samples, the average estimate would be very close to the parameter. There is no systematic error that under/over estimates the statistic.

- If the estimator converges to the parameter on increasing the number of sample points, then the estimator is said to be a **consistent estimator.**

  This implies that using larger samples would result in more accurate estimates. E.g. mean $(sd(\bar{X}) = \frac{\sigma}{\sqrt{n}})$, proportion $(sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}})$ and variance $(sd(S^2_{n-1}) = \frac{\sqrt{2}\sigma^2}{\sqrt{n}})$. It is evident that increasing sample points (n) decreases the spread of the estimation.

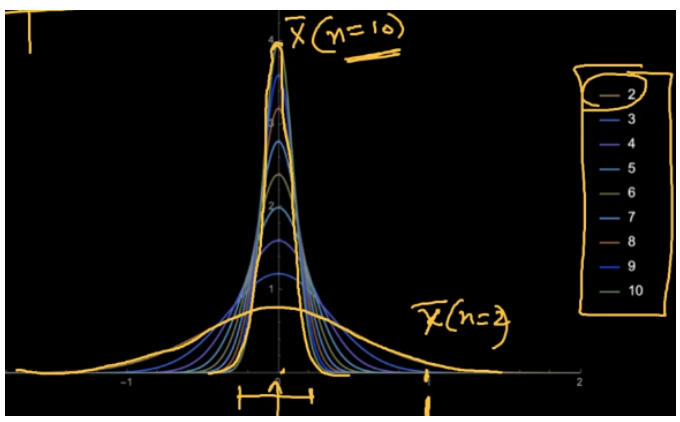


Fig.3 Example of consistent estimator (mean in N(0,1)). As the sample size increases the estimates are closer to the actual mean.

- Estimator optimal with respect to a loss function is said to be an **efficient estimator**. This implies that there should be a mechanism to determine if a new solution is better/worse than the previous one(loss function) and the estimator should minimize this loss function.

Point Estimator for Mean and Proportion

- Sample mean is the point estimate of the population mean. This is plausible as $\bar{X}$ is the unbiased estimator of μ. The standard error is given by $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. The probability of values beyond 2sd from the mean is quite low.

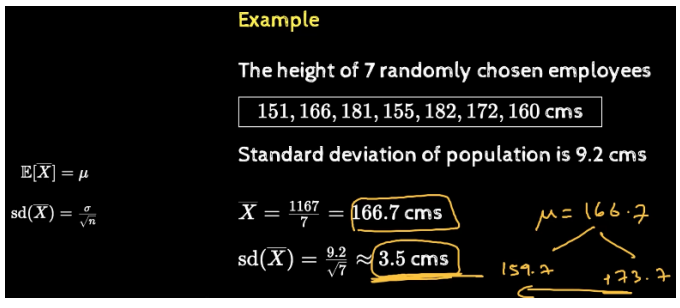


Fig.4a Mean as estimator.

- The sample proportion is the point estimator of the population proportion. The standard error is given by $sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$. Since this requires the population parameter p (which is unknown) the pessimistic value for error is used (the

error is largest at p=0.5). This gives $sd(\hat{p}) = \frac{1}{2\sqrt{n}}$.

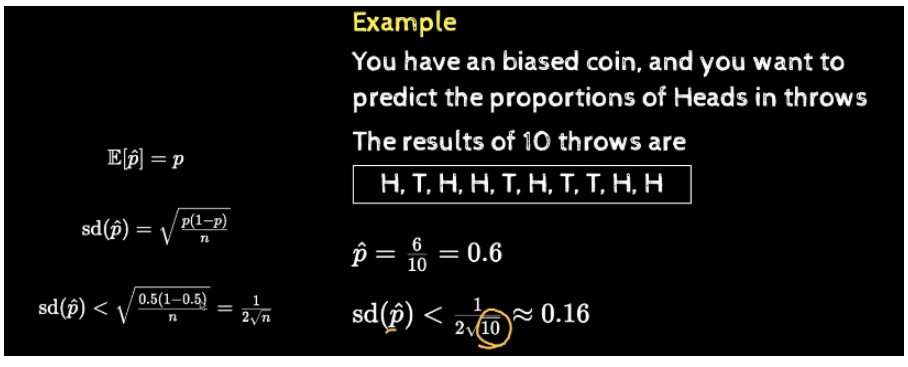


Fig.4b Proportion as estimator.

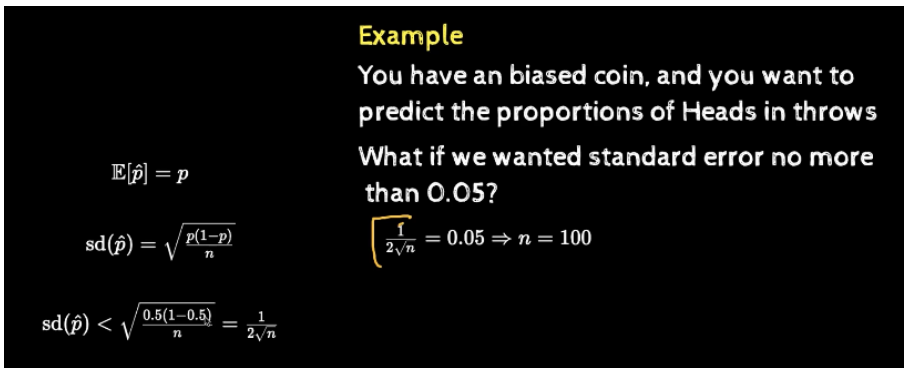


Fig.4c n required for given error margin.

Point Estimator for Sample variance

- The sample variance is used as the point estimator of the population variance. It is an unbiased estimator. The standard error is given by $sd(S^2_{n-1}) = \frac{\sqrt{2}\sigma^2}{\sqrt{n}}$ for normal population distribution.

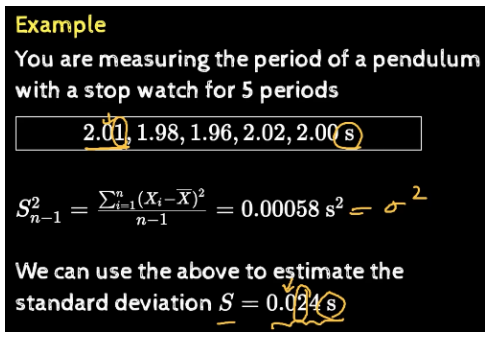


Fig.5a Variance as estimator.

- The standard error of variance is generally not calculated as the error also depends on σ, the parameter being estimated. Also, it works only if the population has a normal distribution.
- If population mean μ is known, then the variance is calculated using μ instead of $\bar{X}$ and the sample size of n is used instead of (n-1) as all n terms are independent of each other.
- The two distributions are scaled versions of the chi squared distribution and differ with respect to the degrees of freedom, where $S^2_{n-1} \sim \chi^2(n-1)$ where as ,$S^2_{\mu} \sim \chi^2(n)$.

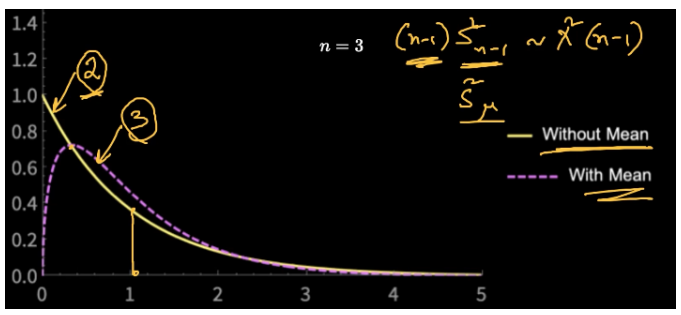


Fig.5b Difference between the distributions.

- The disadvantage of degrees of freedom when μ is not known is small for a large value of n.

Example Estimation with Timeseries

- There is dependency in the timeseries data, but estimation requires that the data points be independent. Thus, instead of studying the raw data, the difference between consecutive time stamps is taken and studied. The difference can be positive or negative, independent of the preceding timestamp.
- A **linear random walk model** is assumed, where the data point at a given timestamp is a combination of the previous timestamp and a random perturbation. It is given by $x_d[n] = x[n+1] - x[n]$, where d is the difference and n is the given timestamp.
- If the linear random walk model is true, it is assumed that x_d **are randomly distributed with a normal distribution.** i.e the positive/negative deviation from the current value is sampled from a normal distribution.

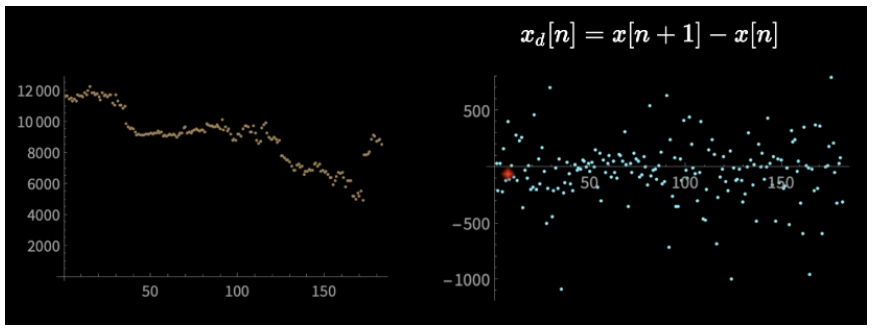


Fig.6a Prices of bitcoin from March-August, 2020. The distribution of difference is random and is centered at 0. The trend seen in the first graph is broken in the second one.

A bell shaped distribution is observed when the distribution of the difference is plotted. Statistics such as variance can be calculated to estimate the population parameters. This is done in the domain of difference which has to be mapped back to the timeseries data.

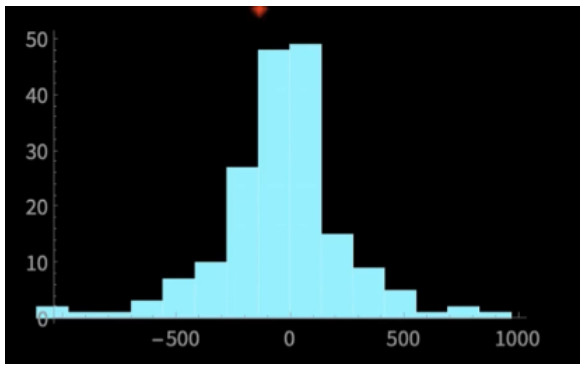


Fig.6b Distribution of the second graph in Fig.6a.

- The differences can also be scaled to calculate relative differences instead of absolute differences using $x_r[n] = \frac{x[n+1]-x[n]}{x[n]}$.
- Here the variable is transformed from one domain to another domain to break the correlation. This is because in statistics the sample values are considered to be independent. The differences are independent of each other and also can be normally distributed (not always).

Real World Problem

- A difference of difference x_{dd} is used if there is a known strong linear trend in the data. $x_{dd}[n] = x_d[n+1] - x_d[n]$.
- If the error in measurements is known, it has to be compensated for in the calculation of statistics(deviation). The measured value = actual vale + error. The **error is independent of the measurement**, then the variance of sum can be written as sum of variance.

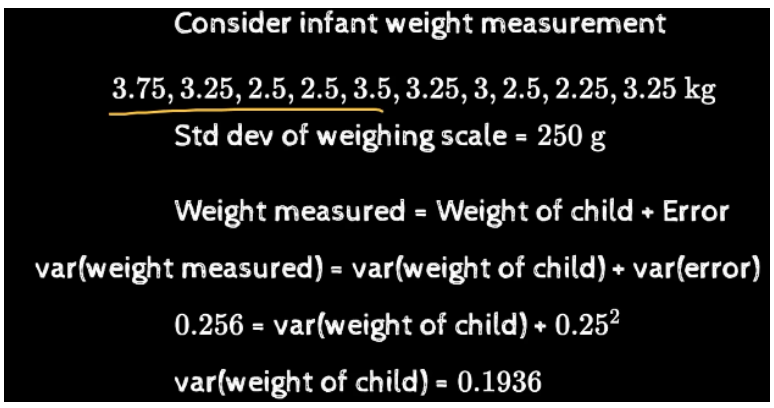


Fig.7 Example of standard error in measurement. It can be observed that the variance is reduced when the standard measurement error is compensated for.

On to Interval Estimators

def interval estimator - (t = 1:00)

def confidence - (t = 1:18)

- Interval estimator **predicts an interval which is expected to contain the parameter.** The confidence of this prediction is the probability that the population parameter is contained within the interval. A confidence of 95% would imply that if a 100 samples were taken then approximately 95 of the samples would contain the parameter in the interval computed for each.
- Interval estimator of population mean cannot be studied for any arbitrary distribution. It holds only for a normal population. This can be over come using a large sample size or a variable made up of sum of independent random variables.

- There can be two cases depending on the availability of the population variance. To estimate the interval of μ with known σ the distribution of $\bar{X}$ should be known .
- The confidence can be calculated using the area under the curve for a given estimate and beyond the interval. It depends on the σ, n and the distribution of the curve.
- The z-score $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is used to write the intervals in terms of μ.

$\Pr(\lvert Z\rvert < 1) \approx 0.68$	$\Pr(\lvert\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\rvert < 1) \approx 0.68$
	$\Pr(\overline{X} - \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}}) \approx 0.68$
$\Pr(\lvert Z\rvert < 2) \approx 0.96$	$\Pr(\overline{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{2\sigma}{\sqrt{n}}) \approx 0.96$

Fig.8 converting interval in terms of Z to interval in terms of mu with corresponding confidence value of 68% and 96%.

Interval Estimator of μ with known σ

- Intervals can be calculated for given confidence p, using $\sigma \; and \bar{X}$ as $Pr(\bar{X} - \frac{I/2}{\sigma/\sqrt{n}} < \mu < \bar{X} + \frac{I/2}{\sigma/\sqrt{n}}) \approx p$. This symmetrically divides the area in the tails to accommodate the given interval.

If the are shaded on one side is u then, the probability p is the unshaded region, $p = 1 - 2u$.

Given p, $u = \frac{(1-p)}{2}$. The I is such that z-table value for $-I/2$ is $(1-p)/2$.

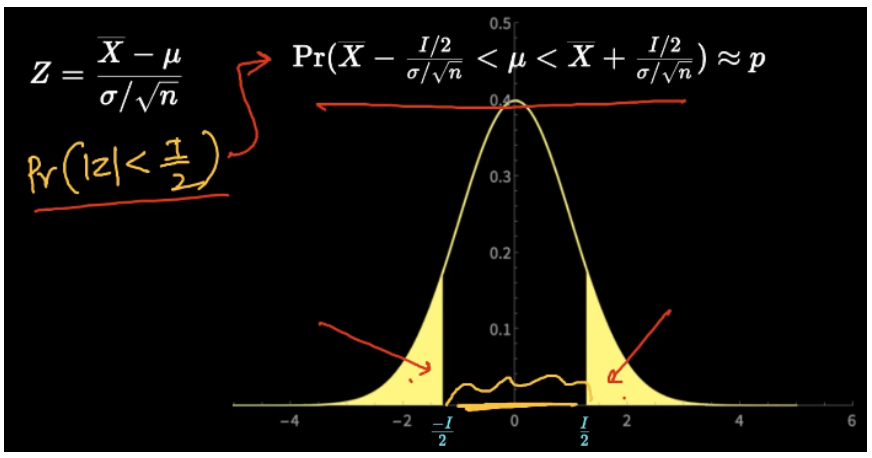


Fig.9a Interval estimator for mean when population deviation is known.

- The relation between $p \; and \; I$ can be made explicit using single variable α and $Pr(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \approx (1-\alpha)$ where, $z_{\alpha/2}$ is such that its z-table score is $\alpha/2$ (the area u = $\alpha/2$), requires inverse lookup of the z-score table. $(1-\alpha)$ is the probability.

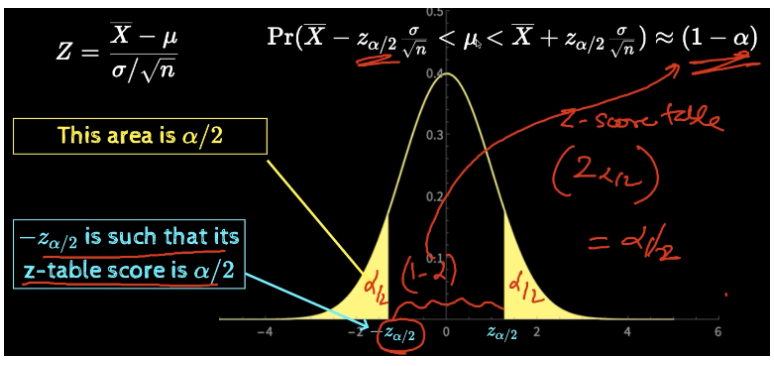
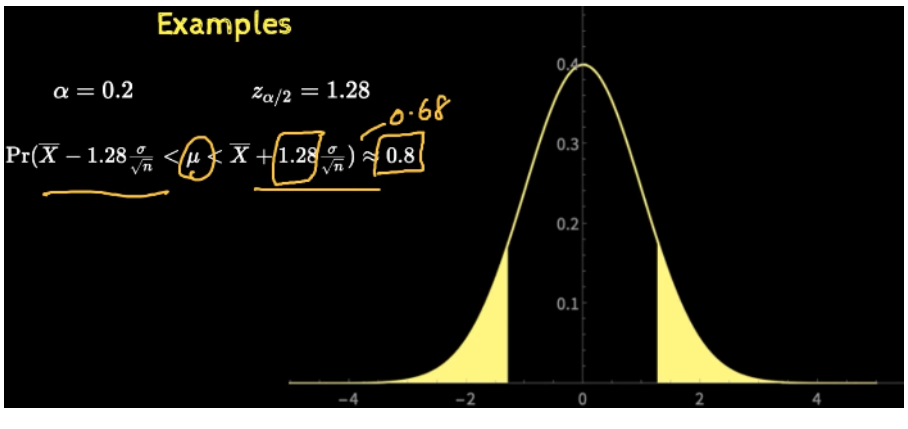
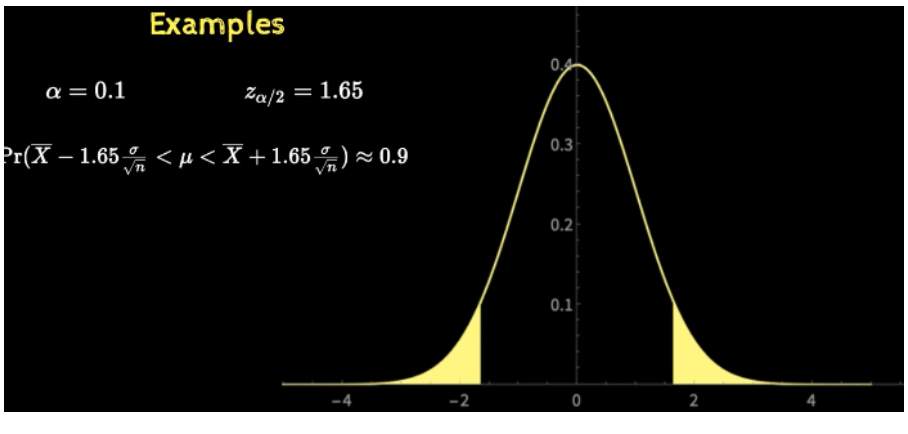
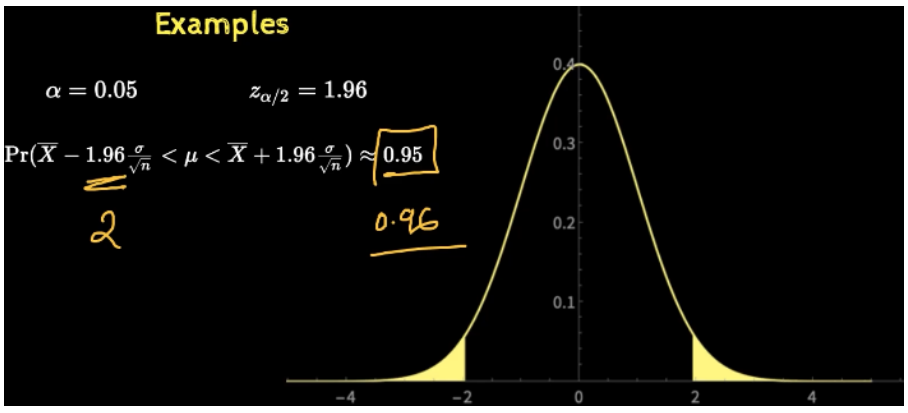


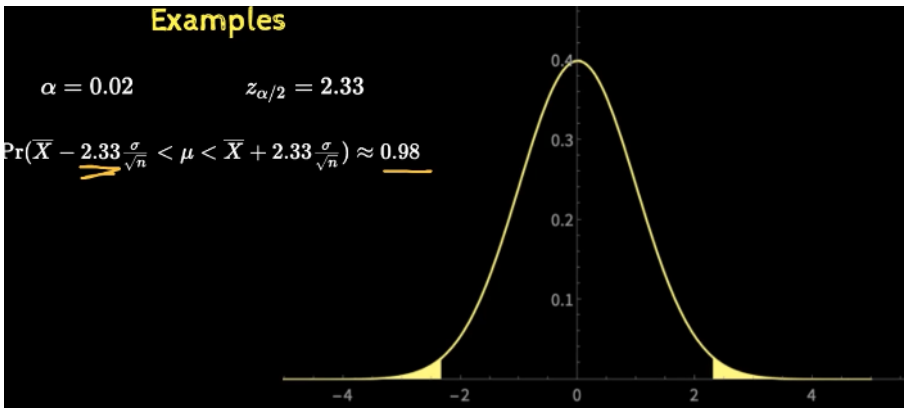
Fig.9b Modified interval estimator for known population deviation.

Examples of Estimator









- It can be observed that as the confidence increases the area under the curve decreases, resulting in wider spread of the confidence interval.

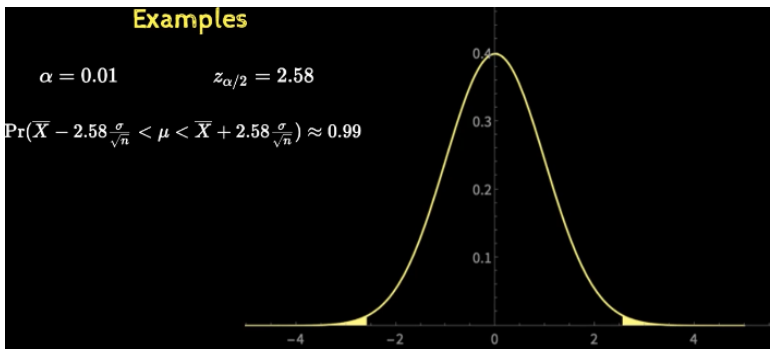


Fig.10 Examples of interval estimator

- For a given μ and confidence $(1-\alpha)$, the interval for different samples is always of the same size as it depends on σ and n and not sample values.
- The interval is plotted around the estimate of $\bar{X}$ that varies from sample to sample. The parameter μ can be found within the interval that $\bar{X}$ occurs with a probability $(1-\alpha)$ and is not found with probability α.

Examples of Estimation

- Thus, μ does not change. $\bar{X}$ changes and the probability is across different samples.
- The example of farmers' market has a known σ since the the relative prices between the shops remains same. All calculations hold under the assumption that the prices are normally distributed.

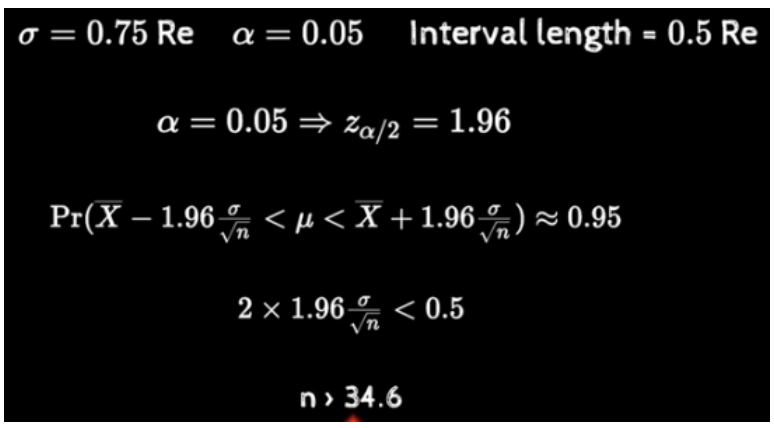


Fig.11 An example of a farmers' market where prices have deviation of 0.75Re and a confidence of 95 is required with in 0.5Re interval length, n has to be calculated.

Lower and Upper Bounds

- Some domains need only the upper/lower bounds of the parameter and not the interval where it may lie e.g. the upper bound of blood pressure.
- While calculating the interval, the area is split into two $(\alpha/2)$, but in upper/lower bound calculation the area is aggregated at one end.
- The lower bound is calculated using $Pr(Z < Z_\alpha) = 1-\alpha$, substituting $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ we get $Pr(\mu > \bar{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-\alpha$. Similarly, the upper bound is given by $Pr(\mu < \bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-\alpha$.
- Calculating lower and upper bounds separately results in tighter bounds than the interval.

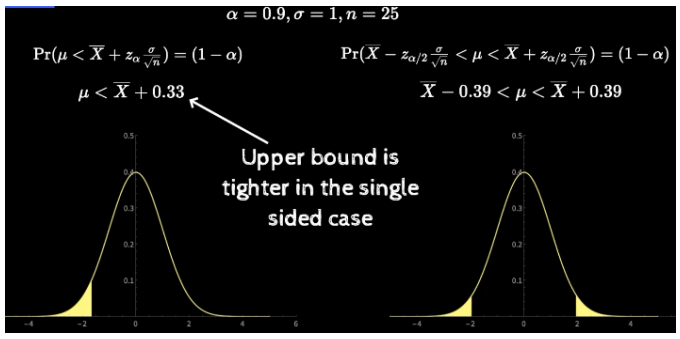


Fig.12 Calculating interval and upper bound for the same confidence.

Upper Confidence Bound

- The upper bound is more conservative estimate with a higher confidence on the calculated parameter.
- It is useful to calculate the bounds (upper/lower) on the mean value to give a sense of possible errors.

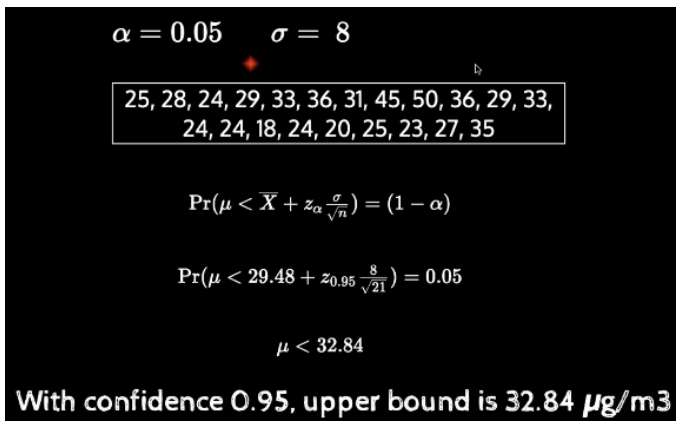


Fig.13 Calculating upper bound on the mean air pollution (pm).

Interval Estimate of μ with Unknown σ

- If the population deviation σ is not known then the sample deviation S_{n-1}can be used instead in $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$, giving $Z = \sqrt{n}\frac{\bar{X}-\mu}{S}$.
- For a normal population or a large n, $Z \sim N(0,1)$. In $\sqrt{n}\frac{\bar{X}-\mu}{S}$, both the numerator and denominator vary with sample $(\bar{X}\ and\ S)$, whereas previously only $\bar{X}$in the numerator varied with sample.
- The following algebraic manipulation is done to accommodate this:

$$T = \sqrt{n}\frac{\bar{X}-\mu}{S}$$

$$T^2 = \frac{(\bar{X}-\mu)^2}{S^2/n} = \frac{(\bar{X}-\mu)^2}{S^2/n}\frac{n/\sigma^2}{n/\sigma^2} = \frac{\frac{(\bar{X}-\mu)^2}{\sigma^2/n}}{\frac{S^2}{\sigma^2}}$$

$$T^2_{n-1} = \frac{Z^2/1}{[(n-1)\frac{S^2}{\sigma^2}]/(n-1)}$$

here, the numerator is a ratio of $\chi^2(1)$ distribution with its degrees of freedom 1 and the denominator is a ratio of $\chi^2(n-1)$ distribution with its degree of freedom $(n-1)$. This ratio is a F-distribution with degrees of freedom 1 and n-1. $T^2_{n-1} \sim F - distribution$ with 1 and n-1 degrees of freedom.

- The numerator will always have 1 degree of freedom, only the denominator will change with the sample size. Thus, $T_{n-1} = \sqrt{n}\frac{\bar{X}-\mu}{S}, T_{n-1}$ is called a t random variable with n-1 degrees of freedom.
- The distribution of T_{n-1} is called the student's t-distribution.

T Distribution Plots

- For n increases S converge to $\sigma$and T_{n-1} looks similar to Z, the standard normal distribution. **For large n, the student's distribution with n-1 degrees of freedom converges to N(0,1).** The difference between the distributions is large for small values of n, i.e. for small sample size.
- Like normal distribution, the student's distribution is symmetric about the y-axis with a mean of 0. But it is shorter than the normal distribution resulting in smaller area in the center. The area is being pushed out to the tails.

- Interval estimate of the mean is done by distributing the uncertainty α among the tails. In student's distribution, since more area is covered in the tails the intervals be larger (due to higher uncertainty in the distribution). The uncertainty is inversely proportional to the sample size.

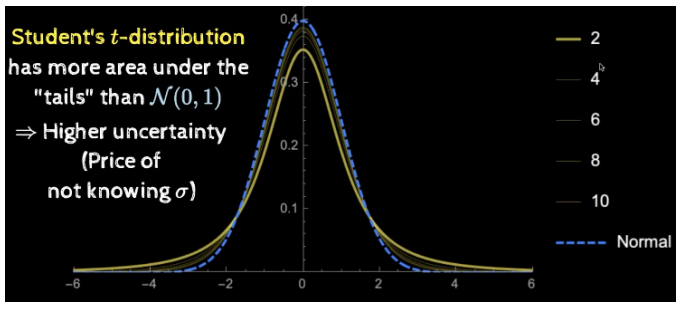


Fig.14 Normal distribution vs student's distribution.

- At n=30 the student's t distribution starts looking similar to the normal distribution but there is still a significant difference. Thus, when σ is unknown it is advisable to use the student's t distribution. At very high degrees of freedom, such as n=1000 the two distributions start looking almost similar.

Comparing Interval Bounds with z- and t-variables

- When σ is unknown the confidence of an interval estimate is given by $Pr(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = (1-\alpha)$. Instead of z-distribution the student's t-distribution is used.

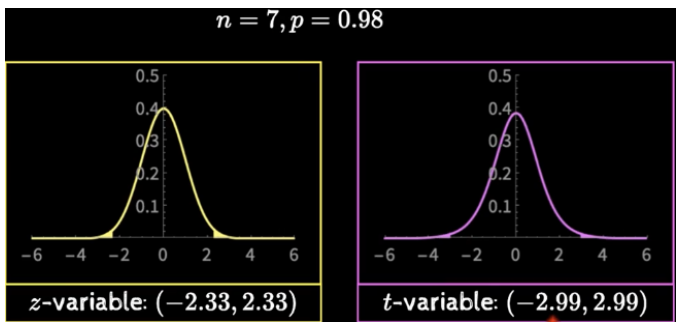


Fig.15a For small n values, a larger area is in the tails of student's distribution, thus, the wider interval than normal distribution for same input.

- Since z is independent of the sample size, it won't change on increasing n, but the student's t-distribution will change . The tails become lighter and the peak becomes taller.

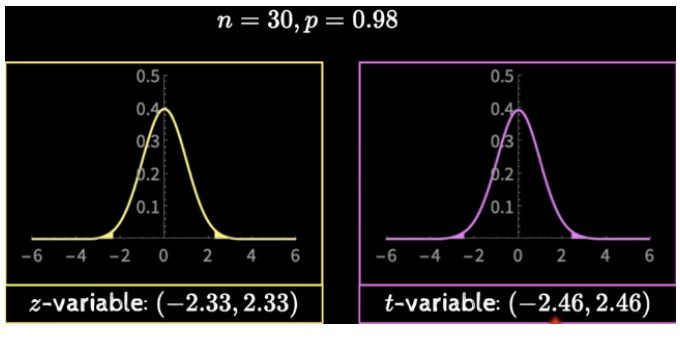


Fig.15b As n increases, the student's distribution approached N(0,1).

- Using t-variables leads to more inaccurate or wider confidence intervals. The difference to z-variables is particularly high for small sample sizes.

Examples with T Statistics

68, 125, 130, 77, 83

$\overline{X} = 96.6 \qquad S = 28.7 \qquad \alpha = 0.05 \qquad n = 5$

$$t_{n-1,\alpha/2} = t_{4,0.025} = 2.78$$

$$\Pr(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = (1-\alpha)$$

$$\Pr(96.6 - 35.7 < \mu < 96.6 + 35.7) = 0.95$$

$$\mu \in (60.9, 132.3)$$

$\overline{X} = 96.6 \qquad \sigma = 25 \qquad \alpha = 0.05 \qquad n = 10$

$$z_{\alpha/2} = 1.96$$

$$\Pr(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = (1-\alpha)$$

$$\Pr(96.6 - 21.9 < \mu < 96.6 + 21.9) = (1-\alpha)$$

$$\mu \in (74.7, 118.5)$$

Fig.16a Example1, the interval is wider when deviation is unknown.

Fig.16b tighter interval when deviation is known .

The bounds for t-statistics maybe narrower for samples where the sample deviation is smaller than σ.

Computing Interval Bounds for Population Proportion p

Recap: The sample proportion is an unbiased estimate of the population proportion $(E[\hat{p}] = p)$ and $sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$. Its a rule of thumb that $n\hat{p}, n(1-\hat{p}) > 10$.

- The assumption that n is large is made such that $sd(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This can be considered as the case with known σ. Thus, to calculate the interval bounds, the likelihood computations with z-statistics can be used (if σ is unknown, t-statistics is used).
- The interval is given by :

$$Pr(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (1-\alpha)$$

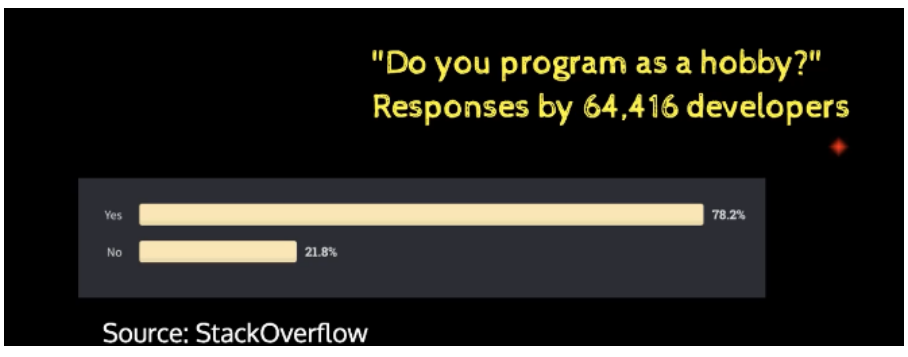
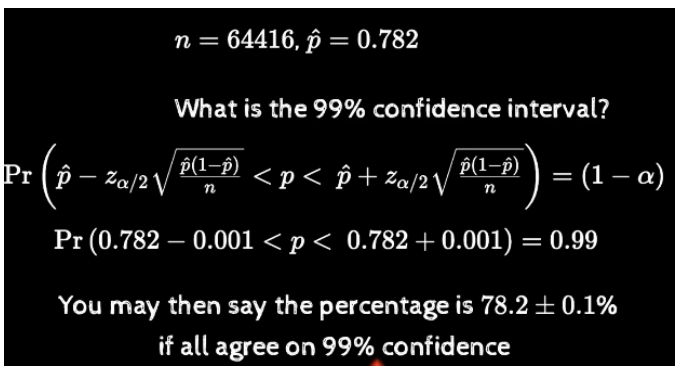


Fig.17 Example of StackOverflow survey.



Summary

- Distribution of statistics over many samples is only studied in theory but in practice, parameters are inferred from a single sample.
- Estimator is a rule that is applied on the sample (e.g. sample mean) to get a concrete value called an estimate, which can be used to predict the population parameter. Based on properties of estimator, it can be unbiased, consistent and efficient.
- While using point estimator mean, population deviation must be known for calculating the standard error. For estimator proportion, pessimistic value (of p = 0.5) is used for calculating the standard error. For variance, the standard error is generally not calculated.
- If there is correlation in the data, it is transformed into different domain where the statistics is done and mapped back to the original domain of interest. Errors due to extraneous factors like the equipment have to be decoupled from the errors in the data.
- For a given sample data, $\bar{X}$ can be calculated and then $\sigma \; and \; n$ can be used to estimate the interval. When σ is known , the confidence is given by $Pr(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \approx (1-\alpha)$.
- Upper bound is calculated using $Pr(\mu < \bar{X} + Z_{\alpha}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ and for lower bound $Pr(\mu > \bar{X} - Z_{\alpha}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ is used. The bounds are tighter than the values calculated using 2 sided bounds.
- For unknown σ, student's t-distribution is used, where $T_{n-1} = \sqrt{n}\frac{\bar{X}-\mu}{S}$. When the degrees of freedom is large, the distribution starts looking similar to N(0,1).
- When σ is unknown the confidence of an interval estimate is given by $Pr(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = (1-\alpha)$. There is more uncertainty in the student's t-distribution which reduces with increase in n.
- For proportions, when $n\hat{p}, n(1-\hat{p}) > 10$, the case of known sigma is assumed and the interval estimate is done using $Pr(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (1-\alpha)$

MCQ

1. A point estimate is
    1. any value from the sample used to estimate a parameter
    2. **a sample statistic used to estimate a parameter**
    3. the margin of error used to estimate a parameter
2. What value will be at the center of a confidence interval
    1. population parameter
    2. **point estimate**
    3. margin of error
3. In 99% confidence interval vs 95% confidence interval
    1. the 95% interval will be wider
    2. **the 99% interval will be wider**
    3. both intervals have the same width
4. Which of the following is not a property of the student's t-distribution?
    1. As the sample size grows, it gradually approaches the normal distribution
    2. Its spread is characterized by the degrees of freedom
    3. It is symmetric
    4. **All of the above are properties of t distribution**
5. The value of α for a 98% confidence interval would be
    1. 0.05
    2. **0.02**
    3. 0.20
    4. 0.10
6. The level of confidence is denoted by
    1. α
    2. **1-α**
    3. β
    4. 1- β
7. σ is unknown, and the sample size n is less than 30, the confidence interval for the population mean is based on the
    1. **t-distribution**
    2. chi square distribution
    3. standard normal distribution
    4. binomial distribution
8. 95% confidence interval for the mean of a population is such that:
    1. It contains 95% of the value in the population
    2. There is a 95% chance that it contains all the values in the population
    3. **There is a 95% chance that it contains the mean of the population**
    4. There is a 95% chance that it contains the standard deviation of the population